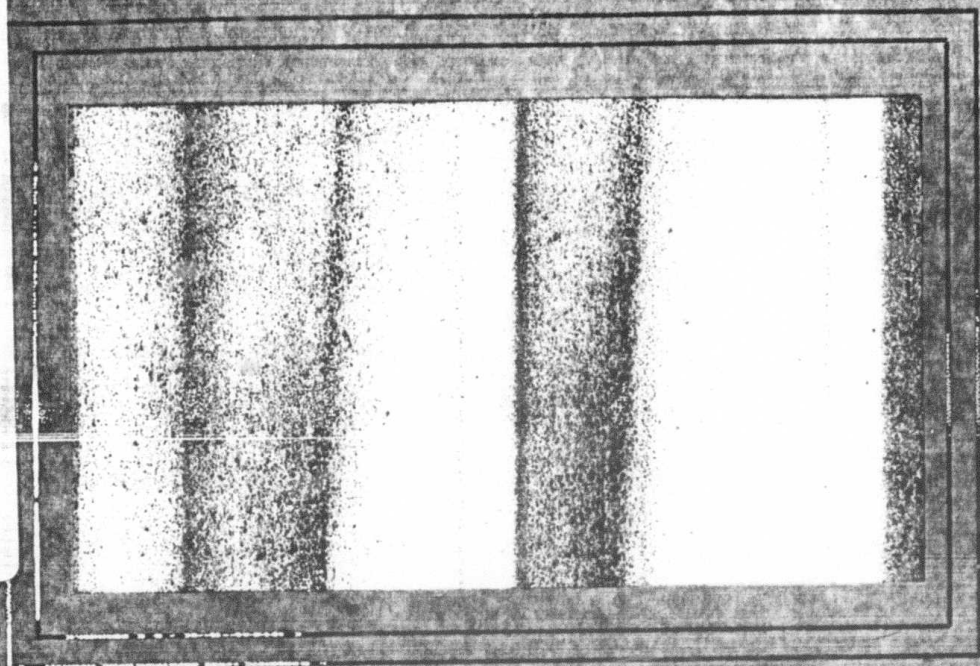


2

AD A124808



DTIC
ELECTE
FEB 24 1983
B

UNIVERSITY OF MARYLAND
COMPUTER SCIENCE CENTER

COLLEGE PARK, MARYLAND
20742

ITC FILE COPY

DISTRIBUTION STATEMENT A

Approved for public release;
Distribution Unlimited

83 02 023 135

TR-1106
DAAG-53-76C-0138

September 1981

BIMEAN CLUSTERING

Stanley Dunn
Ludvik Janos
Azriel Rosenfeld

Computer Vision Laboratory
Computer Science Center
University of Maryland
College Park, MD 20742

ABSTRACT

An algorithm is presented which finds the best-fitting pair of constants, in the least squares sense, to a set of scalar data; we call this pair of constants the "bimean" of the data. The relationship of the bimean clustering to the ISODATA clustering algorithm, and its application to image thresholding, are also discussed.

The support of the Defense Advanced Research Projects Agency and the U.S. Army Night Vision Laboratory under Contract DAAG-53-76C-0138 (DARPA Order 3206) is gratefully acknowledged, as is the help of Janet Salzman in preparing this paper.

DISTRIBUTION STATEMENT A

Approved for public release;
Distribution Unlimited

DTIC
ELECTE
FEB 24 1983
B

1. Introduction

Let us first consider the following minimization problem:

$$\text{minimize } f = \sum_{i=1}^n (x_i - \mu)^2 \quad (1)$$

where we may assume without loss of generality that $x_1 \leq x_2 \leq \dots \leq x_n$. It is well known that f is minimized when μ is the average of the x_i , that is,

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i \quad (2)$$

We now extend this to the following minimization problem:

$$\text{minimize } f = \sum_{i=1}^n (\min((x_i - \mu)^2, (x_i - \nu)^2)) \quad (3)$$

That is to say, we are interested in the best fitting pair of constants to the set x_1, \dots, x_n of data. We refer to this pair of constants (μ, ν) as the bimean. This is because, as we shall see,

$$\begin{aligned} \mu &= \frac{1}{k} \sum_{i=1}^k x_i \equiv \mu_k \\ \nu &= \frac{1}{(n-k)} \sum_{i=k+1}^n x_i \equiv \nu_k \end{aligned} \quad (4)$$

for some k , $1 \leq k \leq n$. That is, the constants μ and ν are such that they are averages or means of subsets of the n data points.

We are interested in the bimean because it defines a natural clustering of the x 's into two subsets. For example, if the x 's are the gray levels of pixels in an image, clustering can be used to segment the image, e.g., into objects and background. Velasco [1] recently showed that the segmentation can be done with the ISODATA clustering algorithm. We will

show that the ISODATA algorithm should ideally converge to the bimean clustering result, but that there are cases where even this ideal clustering approach does not select the appropriate threshold.

This minimization problem has attracted much recent interest. Hartigan and Wong [2] present an algorithm for k-mean clustering which produces a global minimum only for the two-mean case. Pollard [3] discusses the convergence of the k-mean clustering. We consider this question briefly in Section 4 and show that the two-class case of the ISODATA clustering algorithm is relevant to the bimean. Fisher [4] discusses algorithms for clustering, but does not consider the special case of two means.

In Section 2, we develop the bimean clustering algorithm, and in Section 3, we show how the algorithm is applied to image segmentation. Concluding comments are presented in Section 4.



Accession For	
NTIS GRA&I	<input checked="checked" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A	

2. The Bimean Clustering Algorithm

In this section we present an algorithm to compute the values of μ and ν such that

$$f = \sum_{i=1}^n \min((x_i - \mu)^2, (x_i - \nu)^2) \quad (5)$$

is minimized where $x_1 \leq x_2 \leq \dots \leq x_n$. First note that since f is continuous and $f \geq 0$, a minimum exists. We shall show that if among the $n > 1$ values x_1, \dots, x_n there are at least two distinct values, the minimum is attained for some $\mu < \nu$.

Theorem 1: Let $f(\mu) = \sum_{i=1}^k (x_i - \mu)^2$ where $x_1 \leq x_2 \leq \dots \leq x_k$. Then there exists one value of μ at which the minimum is attained, given by

$$\mu^* = \frac{1}{k} \sum_{i=1}^k x_i \quad (6)$$

and for $\mu \neq \mu^*$, $f(\mu) > f(\mu^*)$.

Proof: With μ^* as defined above, we may write

$$f(\mu) = \sum_{i=1}^k (x_i - \mu)^2 \quad (7)$$

$$\text{as} \quad f(\mu) = \sum_{i=1}^k (x_i - \mu^* + \mu^* - \mu)^2 \quad (8)$$

which can be expanded as

$$\begin{aligned} f(\mu) &= \sum_{i=1}^k (x_i - \mu^*)^2 + \sum_{i=1}^k (\mu - \mu^*)^2 + 2(\mu - \mu^*) \sum_{i=1}^k (\mu^* - x_i) \\ &= \sum_{i=1}^k (x_i - \mu^*)^2 + k(\mu - \mu^*)^2 \end{aligned} \quad (9)$$

since the last term is zero with μ^* as defined above. The minimum is obtained when $\mu = \mu^*$, and for $\mu \neq \mu^*$, $f(\mu) > f(\mu^*)$. ||

We now consider the function

$$f(\mu, v) = \sum_{i=1}^n \min((x_i - \mu)^2, (x_i - v)^2) \quad (10)$$

As a corollary to Theorem 1, we now show that the minimum of $f(\mu, v)$ cannot occur for $\mu = v$.

Theorem 2: If the $n > 1$ values $x_1 \leq x_2 \leq \dots \leq x_n$ contain at least two distinct values, the minimum is attained for some $\mu < v$.

Proof: We assume to the contrary that the minimum is attained for $\mu = v$. Theorem 1 implies a minimum at

$$\bar{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \quad (11)$$

Therefore $\bar{\mu} < x_n$, because x_1, \dots, x_n contain at least two distinct values. We now consider

$$f(\bar{\mu}, x_n) = \sum_{i=1}^n \min((x_i - \bar{\mu})^2, (x_i - x_n)^2) \quad (12)$$

and we see that

$$f(\bar{\mu}, x_n) \leq \sum_{i=1}^{n-1} (x_i - \bar{\mu})^2$$

and also that

$$\sum_{i=1}^{n-1} (x_i - \bar{\mu})^2 < \sum_{i=1}^n (x_i - \bar{\mu})^2$$

since $x_n - \bar{\mu} > 0$.

Since $f(\bar{\mu}, \bar{\mu}) = \sum_{i=1}^n (x_i - \bar{\mu})^2$ and from Theorem 1 follows that

this expression attains the unique minimum at $(\bar{\mu}, \bar{\mu})$.

We see that

$$f(\bar{\mu}, x_n) < f(\bar{\mu}, \bar{\mu})$$

which is a contradiction to our assumption that $f(\mu, v)$ is

minimized when $\mu = v$. We conclude that $\mu \neq v$ if the $n > 1$ values $x_1 \leq x_2 \leq \dots \leq x_n$ contain at least two distinct values. ||

From the very definition, it follows that $f(\mu, v) = f(v, \mu)$. With this fact and Theorem 2, we can restrict the domain of definition of the function $f(\mu, v)$ to be $\{(\mu, v) : \mu, v \in \mathbb{R} \text{ and } \mu < v\}$.

We now define the numbers μ_k and v_k to be

$$\begin{aligned}\mu_k &= \frac{1}{k} \sum_{i=1}^k x_i \\ v_k &= \frac{1}{n-k} \sum_{i=k+1}^n x_i\end{aligned}\tag{13}$$

for all k , $1 \leq k \leq n$. We prove our final theorem from which the bimean clustering algorithm follows.

Theorem 3: If (μ^*, v^*) is the minimum of $f(\mu, v)$ for $\mu < v$ then there exists an index k such that

$$\begin{aligned}\mu^* &= \mu_k, \quad v^* = v_k \\ x_k &\leq \frac{1}{2}(\mu^* + v^*) \\ x_{k+1} &> \frac{1}{2}(\mu^* + v^*)\end{aligned}\tag{14}$$

Proof: Let k be the largest index such that

$$x_k \leq \frac{1}{2}(\mu^* + v^*)$$

is true. We can write $f(\mu^*, v^*)$ as

$$f(\mu^*, v^*) = \sum_{i=1}^k (x_i - \mu^*)^2 + \sum_{i=k+1}^n (x_i - v^*)^2\tag{15}$$

Assume to the contrary that $\mu^* \neq \mu_k$ or $v^* \neq v_k$

Case 1: $\mu^* \neq \mu_k$.

Consider $f(\mu_k, v^*) = \sum_{i=1}^n \min((x_i - \mu_k)^2, (x_i - v^*)^2)$

$$\leq \sum_{i=1}^k (x_i - \mu_k)^2 + \sum_{i=k+1}^n (x_i - v^*)^2 \quad (16)$$

By Theorem 1, the first term of equation (16) is minimized, thus

$$f(\mu^*, v^*) > \sum_{i=1}^k (x_i - \mu_k)^2 + \sum_{i=k+1}^n (x_i - v^*)^2 \quad (17)$$

From equations (16) and (17)

$$f(\mu_k, v^*) < f(\mu^*, v^*)$$

and we have achieved a contradiction to the assumption that the minimum of $f(\mu, v)$ is attained at (μ^*, v^*) .

Case 2: $v^* \neq v_k$.

The proof for this case is analogous to that for the first case and shall be omitted. Thus we have shown that $\mu^* = \mu_k$ and $v^* = v_k$ for some k , $1 \leq k \leq n$. ||

With Theorem 3 in hand, we present the bimean clustering algorithm:

Step 1) Find the set K of indices which satisfy

$$\begin{aligned} x_k &\leq \frac{1}{2}(\mu_k + v_k) \\ x_{k+1} &> \frac{1}{2}(\mu_k + v_k) \end{aligned}$$

Step 2) For all $k \in K$ evaluate $f(\mu_k, v_k)$

Step 3) Find the minimum of $f(\mu_k, v_k)$ for all $k \in K$. Set $\mu^* = \mu_j$ and $v^* = v_j$ where j is the largest element of K for which the minimum is attained.

3. Application to thresholding

A possible application of the bimean is to segment an image so as to separate objects from their background, by clustering the gray levels of the image's histogram into two clusters. Thus, the index k in the bimean clustering algorithm is the gray level above which the gray levels belong to the objects, and the cluster of gray levels below the index k belong to the background.

Figure 1 shows results of applying the bimean clustering algorithm to a set of infrared images of tanks. The original images are shown in the first column, and the bimean results in the third column. The second column shows ISODATA results (see below), with the initial threshold taken at the mean gray level of the image. We see that the first five images are reasonably segmented by the bimean algorithm, but the last three are not; and the ISODATA results are not as good (e.g., the fourth image is poorly segmented).

Velasco [1] showed that a two-class, one-dimensional ISODATA clustering algorithm (e.g., [5]) could be used to segment images into two gray level classes. Our experiments show that this is not always the case. ISODATA is an iterative process based on the same distance measure that we are minimizing in the bimean algorithm; but ISODATA may converge only to a local minimum of this measure, whereas our algorithm finds the global minimum. In spite of this, we do not always obtain good thresholds, and neither does ISODATA.

4. Discussion and concluding remarks

We have presented a new algorithm for clustering single dimensional data into two clusters. The computation is relatively quick, and the equations can be rewritten so as to only perform a single pass through the data.

The ISODATA algorithm should ideally converge to the Bimean, but this requires a suitable initial choice of means for ISODATA. For example, if the second mean is set equal to one outlier, the ISODATA algorithm converges but possibly not to the true bimean. Thus, the Bimean algorithm is a more reliable method of obtaining a globally optimal threshold than iterative algorithms such as ISODATA. Figure 1 shows, however, that this method does not always perform well in practice.

References

1. Velasco, F.R.D., "Thresholding using the ISODATA Clustering Algorithm," IEEE Transactions on Systems, Man, and Cybernetics, vol. 10, 1980, pp. 771-774.
2. Hartigan, J. A. and M. A. Wong, "A K-Means Clustering Algorithm," Applied Statistics, vol. 28, 1980, pp. 100-108.
3. Pollard, D., "Strong Consistency of K-Means Clustering," Annals of Statistics, vol. 9, 1981, pp. 135-140.
4. Fisher, W. D., "On Grouping for Maximum Homogeneity," Journal of the American Statistical Association, vol. 53, 1958, pp. 789-798.
5. Duda, R. D., and P. E. Hart, Pattern Classification and Scene Analysis, John Wiley & Sons, New York, 1973.

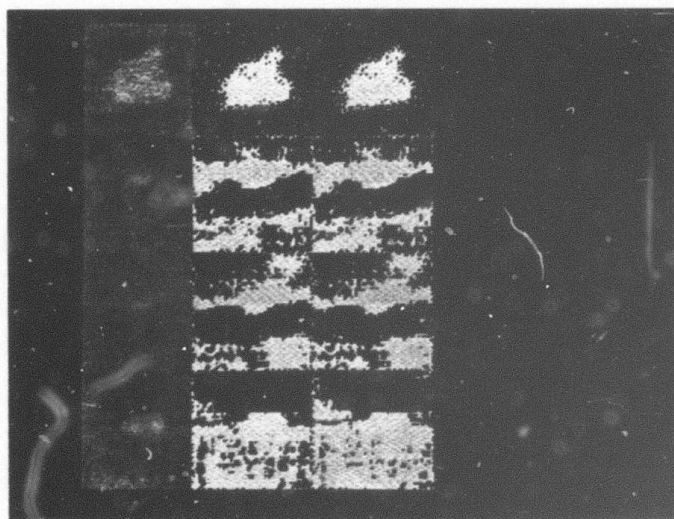
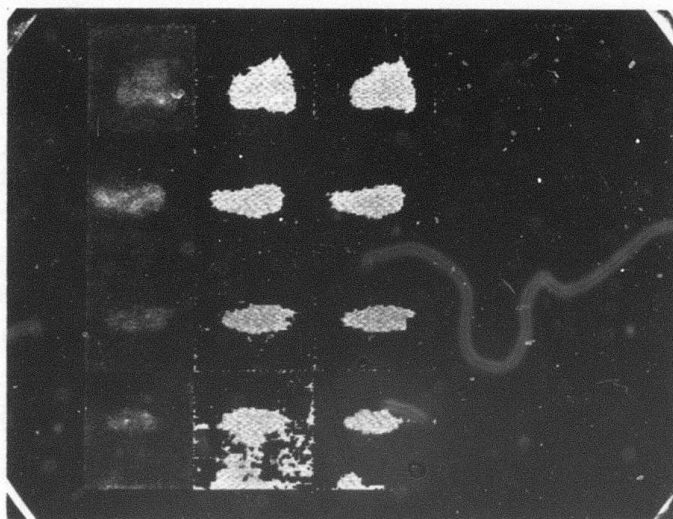


Figure 1

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO. ADA124808	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) BIMEAN CLUSTERING		5. TYPE OF REPORT & PERIOD COVERED Technical
		6. PERFORMING ORG. REPORT NUMBER TR-1106
7. AUTHOR(s) Stanley Dunn Ludvik Janos Azriel Rosenfeld		8. CONTRACT OR GRANT NUMBER(s) DAAG-53-76C-0138
9. PERFORMING ORGANIZATION NAME AND ADDRESS Computer Vision Laboratory Computer Science Center University of Maryland College Park, MD 20742		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS U.S. Army Night Vision Laboratory Ft. Belvoir, VA 22060		12. REPORT DATE September 1981
		13. NUMBER OF PAGES 10
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Image processing Pattern recognition Clustering Segmentation Thresholding		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) An algorithm is presented which finds the best-fitting pair of constants, in the least squares sense, to a set of scalar data; we call this pair of constants the "bimean" of the data. The relationship of the bimean clustering to the ISO/ATA clustering algorithm, and its application to image thresholding, are also discussed.		

DD FORM 1473

1 JAN 73

EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)